ISSN No. 0976-5697

Volume 5, No. 8, Nov-Dec 2014

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Lip Motion based Alphabet Recognition using Neural Network

Disha George
Dept. of Computer Science and Engineering
Raipur Institute of Technology
Raipur, India

Yogesh Rathore
Dept. of Computer Science and Engineering
Raipur Institute of Technology
Raipur, India

Abstract: The facial expressions in image sequence deliver information in context to the subject spoken by the speaker but it adds a challenge when it to be represented in animation system. It has great impact in the field of audio-visual speech recognition (AVSR). Some of us are capable of Lip reading by interpreting the lip motion. This research is divided into two-levels: (i) Firstly, frames are taken and features are extracted from these frames to be kept in database as standard. (ii) Secondly, test image samples to be trained in neural network to check the alphabet spoken with the trained images (stored in database) to recognize what the person has spoke. Lip reading system has been developed using K- Means Clustering using input images and 60% success has been achieved in similar alphabets lip movements (such as u, o, q, b, e, i, l, n etc.).

Keywords: Lip Motion, K-Means Clustering, Pattern Recognition, Artificial Neural Network

INTRODUCTION

Synthesizing Lip movements and automatic understanding is a tedious task. Lip reading of particular subject has carried out through years for dumb and deaf people to recognize what the person has spoken and communicate in an effective way [1]. Visual Signs plays as supplementary role of audio signals to render complete information [13]. Lip reading is examined in two perspectives: speech recognition and analysis of visual signs [8]. Systems based on audio signals are not fully reliable since they get affected drastically due to noises. Visual information gives only 1/3 of the conveyed message [2], [11], [12]. In this research, the framework is divided into two as: (i) Pre-processing: the video corpus is taken and converted into sequence of frames and the *feature extraction* is done and all these are gathered into database. (ii) Post-processing: After performing the step (i), training of test images are to find matching between the contexts. The comprehension of speech in the lack of sound is called as lip reading. A framework is presented that tackles the images of lip portion and is capable of synthesizing and analyzing.

The images are selected according to the variations from the number of frames [1]. Accordingly, the features such as inner width, outer width, height between outer lips, height between inner lips, distance between dip and peaks of lips are extracted from the images [3],[13]. These features along with images are collected in the database [5]. The same is applied to the set of test images and are trained in neural network to find the best proximity. Feature vector is collected from the lip shape in the database using *approximating with K-Means Clustering*. This database is used to find the closest proximity between the base and test images of particular subject spoken by the person [9]. *Artificial Neural Network* (ANN) is used for training function for best matched lip shape received by lip motion while pronouncing particular alphabet to be identified [18].

PROPOSED METHOD

Lip reading system comprises an operation which aids to understand what the speaker has spoken without the requirement of audio message. It contains complex computational methods. The proposed system is subdivided into sub units and each sub unit is processed and analysed separately to collect every bit of information deeply.

In the proposed model, firstly the sequence of images is taken and then the lip portion is carved as a part of work to be processed [14]. Here K-Means clustering is used so as to get the lip shape from the image [16]. Later, points in the contour of the lips are gathered from the frame to form the feature vector matrix [17]. A feature matrix is drawn out by performing operations on shapes of lips and feature vector extraction for the following frames of various alphabets spoken by speaker [6]. This feature matrix is inserted into the database by applying with segmentation technique [15]. When a test sample image is reviewed, then the similar procedure is applied and is compared with the database maintained of the frames [9]. The feature extraction is done and this vector of test image is compared to database to find the close proximity with any of the image feature vector in the database so as to find what the speaker says[18]. The steps of operation are applied can be seen further.

A thresholding technique is applied to determine the lip region [7].Binary lips are obtained after transformation by strengthening the lips. K-Means clustering segmentation method is applied so as to get the particular lip shape of interest from image. This benefits to detect the lip area and give a particular shape of the determined lip shape. Morphological operations are applied so as to obtain the shape and contours. Feature Vector is obtained from the binary image [10].

The complex work to be done in lip-reading is the feature extraction of the supposed lip image [4]. The coordinates are used for features to be extracted. The attributes to be worked over are *the height between the inner* *lips, the height of outer lips, visibility of teeth in between, the corners of the lips, width of edges of lip, the coordinates representing the contour of lips, distance between dip point and peaks of lip, height of lower inner and upper outer lip*. The feature vector is prepared by finding these all assets one by one. The whole procedure can be carried out in 3 levels. This can be shown in the figure below:-

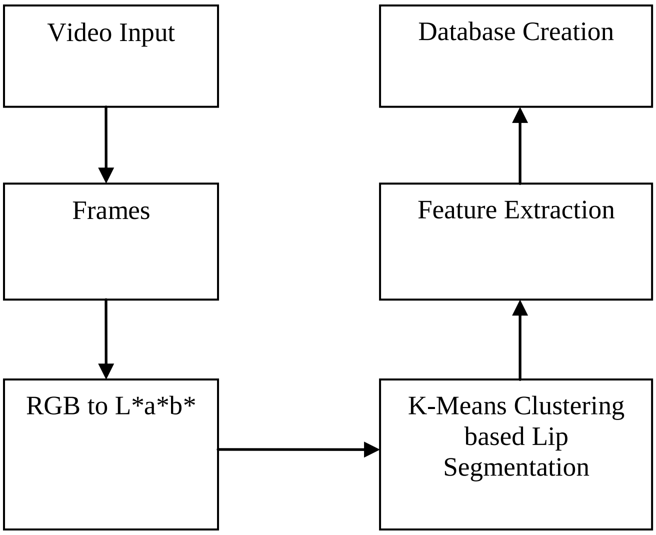

Fig.1 The block diagram of Database Creation

A. *Lip Motion tracking system:*

(I) Conversion of RGB Color Space to L*a*b*:

The images are converted from RGB to L*a*b color space. The image is read. The L*a*b* color space known as CIELAB or CIE L*a*b* enables to quantify these visual differences. It is originated from CIEXYZ tristimulus values. The L*a*b* space comprises of a luminosity layer 'L*', chromaticity-layer 'a*' indicating where does the color falls along the red-green axis, and chromaticity-layer 'b*' indicating where does the color falls along the blue-yellow axis. All of the color aspects lies in the 'a*' and 'b*' layers. The difference between two colors is measured using the Euclidean distance metric.

(II) Classify the colors using K-Means Clustering:

Clustering is a technique which partitions the objects into groups of different clusters. K-Means clustering partitions each object within cluster are as close to each other or how far are from other objects in the cluster .It is necessary to specify the number of clusters required to be partitioned and the distance metric to evaluate the distance between the objects within the clusters. However, the color information is in a*b* space so the objects are pixels with a* and b* values.

K-Means is one of the most popular clustering algorithms, which produces non-overlapping clusters. Individual cluster has a centroid (also known as seed), which represents the general features of the cluster.

$$W(S,C) = \sum_{k=1}^{K} \sum_{i \in S_k} d(i\ ,C_k) \qquad (1)$$

In Eq. 1, The General K-means criterion, d is the squared Euclidean distance. It is as:

$$W(S\ ,C) = \sum_{k=1}^{K} \sum_{i\ \in\ I} \sum_{v=1}^{M}\ S_{ik}(Y_{iv} - C_{kv})^2 \qquad (2)$$

Eq. 2, depicts K-means criterion using the squared Euclidean distance. In the above Eq. 2, S_{ik} is a Boolean variable representing the cluster membership of i to the cluster k. Formally, will be equal to one if and only if $i\ \in\ S_k$.

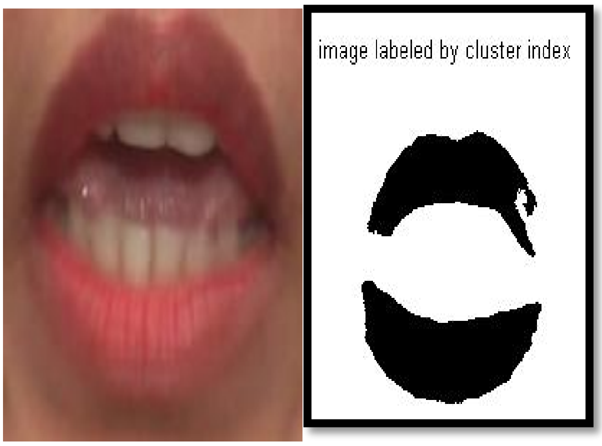

Fig.2. (a) Original Image (b) Segmentation by K-Means

III) Creation of Smallest Rectangular Window:

The region of interest is the pair of lips in the image and so is focused by covering it in Smallest Rectangular window. In the sequence video, the Smallest Rectangle Window (SRW) is detected on an input image, it takes few steps:

(i) To set the rectangle window so as to cover the lip motion sufficiently by excluding the surrounding parts as nose or other face portion unnecessarily. The size of new rectangle in consecutive image is taken double of the current SRW of image.

(ii) The detection is done on pixels of skin after segmenting the normalized image into skin and non-skin areas using thresholding. The use of thresholding contains two clear peaks (lips or skin around lips) and one dip or valley, the peak is small for mouth open. Segmentation is performed by peak and valley thresholding variant.

(iii) Extraction of largest white or black connected surface and determination of SRW containing lips on succeeding image.

The Region of Interest containing the mouth is fixed explicitly in the first image in the sequence video. About nine or ten features describing the geometry and texture of lip are considered based on height, width and their variants.

III) Feature Extraction of Lips:

The feature extraction is then done after the above steps so as to obtain the properties by which the alphabet may be recognized by lip shape. The various features which are evaluated are as follows (some of them are shown in figure):

(a) Major axis (1).

(b) Height of Minor Axis (2).
(c) Width of lips (3).
(d) Height between upper inner lip and lower outer lip (4).
(e) Height between lower inner lip and upper outer lip (5).
(f) Distance between valley and peaks (6).
(g) Area of lips' portion.
(h) Perimeter of lips.
(i) Distance from valley of lip to midpoint of lower outer lip.

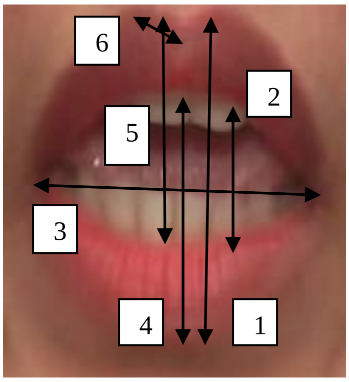

Fig.3 Representation of feature dimensions

These dimensions of lips are retained and a *feature matrix* is formed. This feature of each frame is gathered and integrated for each frame in the database for different frames of various subjects. Distinct alphabets are spoken by the speaker and are collected in the database so that these can be used further for training purpose.

B. *Training Phase:*

Artificial Neural Network is the computing system inspired by human brain is a collection of simple, highly connected elements which process information according to their dynamic state response to external inputs in order to generate outputs. Neural Network is organized in layers which are made up of interconnected nodes containing activation function. Patterns are introduced to the network through the *input layer* which corresponds to one or more *hidden layers* where the actual processing is done via system of *weighted connections.* It is applied for pattern recognition, data classification through learning.

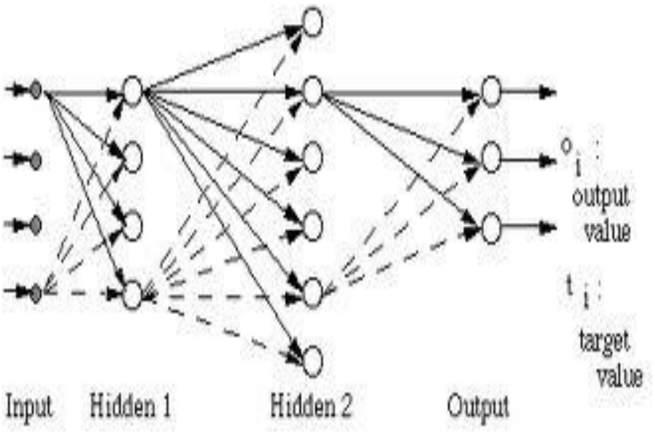

Fig. 4 Artificial Neural Network

A radial basis function network is a network like an artificial neural network. They accept numerical inputs, and generate a number of numerical outputs and can be used to make calculations. In training, RBFN is to determine the set of weights and bias values that build a network whose outputs best match with those of the training data. Training a radial basis function network comprises three important steps. In the key step, a set of centroids is found, a centroid for each hidden node. In the second step, a set of widths is definite, individual width value for every hidden node. In the third step, a set of weights is calculated. A set of attributes or features, as mentioned earlier in the paper is used to train to find the similar pattern with the testing sample and if they matches, then that will be the best match and considered output. Radial Basis Function Network (RBFN) is an proficient solution for the researchers who are working on the field of machine recognition, pattern recognition and computer vision. The main challenge in the face recognition technology is to deliver high recognition rate. In this paper, a helpful method has been presented for pattern recognition using K-Means clustering and radial basis function. To be specific, segmentation has been used for feature extraction of the dataset and radial basis function network has been used as a classifier for classification of data and for recognition process also.

In this phase, the database with feature matrix of different subjects is having 26 alphabets of 10 speakers and 20 frames of each alphabet. During training phase, the network is trained to relate output with input patterns. The training phase can be demonstrated as in the following figure:

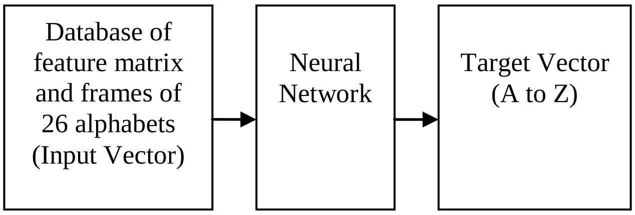
Database of feature matrix and frames of 26 alphabets (Input Vector)
Neural Network
Target Vector (A to Z)

Fig.4 Block diagram of training phase

Here, Input Vector from database is given as input to the neural network and after processing in neural network, the target vector of 26 alphabets is generated so as to benefit in testing phase.

C. *Testing Phase:*

In this phase, the unknown Input Vector is taken as input. In testing phase, the network output identifies output from training phase with the associated output. In such condition, the network renders output that corresponds to a taught input pattern that is least different from given the pattern (in case of recognition of alphabets or symbols.). It is then passed to Neural Network to go through the predefined process. Later, the output is generated where the matched properties depicts the alphabet spoken by speaker. The matched helps to recognize the alphabet by the lip shape as seen in the image. It can be shown in the figure below:

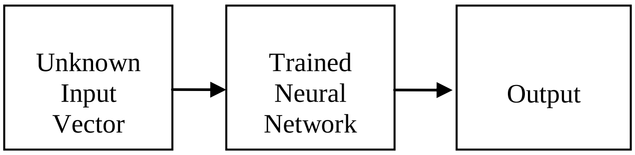

Fig. 5 Block diagram of testing phase

The Neural Network is trained one. The output generated shows the result as the individual which matched with the alphabet spoken as given in the base sample. It signifies the alphabet which is recognized by the lip shape.

III Result

In this paper, we have proposed a method by which we can find the alphabet spoken recognising the lip shape of the speaker. Here, we have used the Radial Basis Function Network (RBFN) for the training and pattern recognition purpose. The video corpus is taken as input and is collected as frames; these frames are segmented to find features. These features are combined so as to form feature matrix. The paper delivers better approximation between the patterns identified. In training stage, we experienced with some patterns.

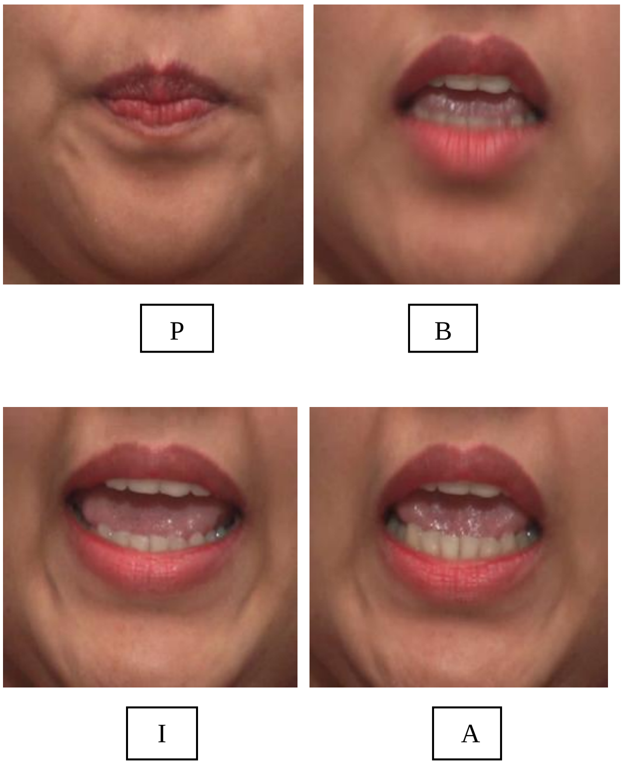

Fig. 6 Frames taken from videos of particular subjects

Table II RESULTS OF SOME ALPHABETS

Alphabets	**Recognition Results**		
	First nearest	Second nearest	Third nearest
A	I (65%)	İ (20%)	A(80%)
E	E (50%)	I (70%)	B(65%)
B	B (85%)	E (30%)	I (35%)
H	H (82%)	J (15%)	G (5%)
O	U (64%)	I (25%)	Q (35%)
S	S (75%)	X (35%)	Z (15%)
U	U (87%)	I (50%)	O (60%)

Despite of this, we achieved 65% success in the work of alphabet recognition based on lip motion. Some of the information is conveyed by audio signal and some by visual. It is effective if visual information is taken as auxiliary aid to audio which gives more successful result in human interaction with computer system.

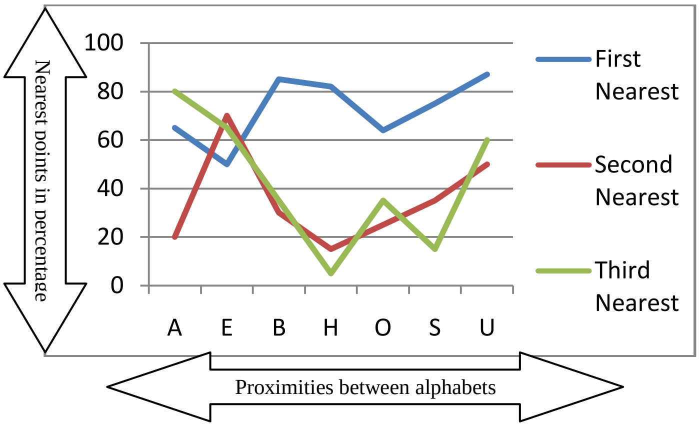

Fig.7 Graph showing closest similarities between alphabets

The graph illustrates the relation between the proximities of alphabet and the similarities among the alphabets falling close by to the test sample of alphabet. It is seen from the graph that similarities can be found between the nearby lip shapes of various subjects but the alphabet having the most proximity will be the exact one.

CONCLUSION

The research is based on the lip shape based on lip motion. The segmentation technique that is K-Means clustering helps to find the features and Radial Basis Function Network of Neural Network benefits for pattern recognition of lip shape which defines the identification of alphabet spoken by the speaker. The further work can be carried on by applying DCT, ACM Model or on the basis Bezier curves.

References

[1] Abhay Bagai, Harsh Gandhi, Rahul Goyal,Ms. Maitrei Kohli, Dr. T.V.Prasad," Lip-Reading using Neural Networks" IJCSNS International Journal of Computer Science and Network 108 Security, VOL.9 No.4, April.

[2] A.Rogozan, P.Deléglise, Visible Speech Modeling and Hybrid Hidden Markov Models / Neural Networks Based Learning for Lipreading. IEEE Computer Society, France (1998).

[3] I. Matthews, T.Cootes, J.Bangham, S.Cox, R. Harvey, Extraction of Visual Features for Lip reading. PA&MI, IEEE Transaction, 198-213, (2002).

[4] I. S. Lindsay. A Tutorial on Principal Components Analysis, USA, (2002).

[5] J. Ma, J. Yan, and R. Cole, "CU Animate: Tools for Enabling Conversions with Animated Characters," Proc. Int'l Conf. Spoken Language Processing, pp. 197-200, 2002.

[6] R. Cole, S. Van Vuuren, B. Pellom, K. Hacioglu, J. Ma, J. Movellan,S. Schwartz, D. Wade-Stein, W. Ward, and J. Yan, "Perceptive Animated Interfaces: First Steps toward a New Paradigm for Human-Computer Interaction," Proc. IEEE, vol. 91, no. 9, pp. 1391-1405, 2003.

[7] G. Geiger, T. Ezzat, and T. Poggio, "Perceptual Evaluation of Video-Realistic Speech," CBCL Paper #224/AI Memo #2003-003, Mass. Inst. of Technology, Cambridge, Mass., Feb. 2003.

[8] S.W. Choi, D. Lee, J.H. Park, and I.B. Lee, "Nonlinear Regression Using RBFN with Linear Sub Models," Chemometrics and Intelligent Laboratory Systems, vol. 65, no. 2, pp. 191-208, 2003.

[9] J. Ma and R. Cole, "Animating Visible Speech and Facial Expressions," The Visual Computer, vol. 20, nos. 2-3, pp. 86-105, 2004.

[10] S.Wamg, H.Lau, S.Leng, H.Yan. A Real Time Automatic Lip reading System. ISCAS'04, vol:2, p.101-104, Hong Kong, (2004).

[11] L.Xie, X.Cai, Z.Fu, R. Zhoa, D.Jiang. A Robust Hierarchical Lip Tracking Approach for Lip reading and Audio Visual Speech Recognition. Proceedings of the 3rd ICMLC, 3620-3624, Shangai, China, (2004).

[12] V.V. Nabiyev, Artificial Intelligence: Problems-Methods-Algorithms (in Turkish), Seckin Publishing, 2nd Press, (2005).

[13] Zafer Yavuz, Vasif Nabiyev **,"**AUTOMATIC LIPREADING WITH PRINCIPLE COMPONENT ANALYSIS", supported by Karadeniz Technical University Research and development fund for the science project N.2005,112.009.01.

[14] Jiyong Ma, Member, IEEE, Ron Cole, Member, IEEE, Bryan Pellom, Member, IEEE,Wayne Ward, and Barbara Wise," Accurate Visible Speech Synthesis Based on Concatenating Variable Length Motion Capture Data", IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 12, NO. 2, MARCH/APRIL 2006.

[15] Z.Yavuz, V.V.Nabiyev. Automatic Lipreading, 15th SIU'07, Eskişehir, (2007).

[16] A. Rouigueb, S. Chitroub, and A. Bouridane, Senior, *IEEE,"* Fuzzy Local ICA for Speaker Recognition Using Voice and Lip Motion, Proceedings of the World Congress on Engineering 2012 Vol I WCE 2012, July 4 - 6, 2012, London, U.K.

[17] Sunil S.Morade, Suparva Patnaik , "Visual Speech Recognition using Features of Lip and Effect of Database on Digit Recognition" , Proc. of the Intl. Conf. on Advances in Computer, Electronics and Electrical Engineering, pp. 456-459, 2012.

[18] Disha George, Yogesh Rathore, "LIP MOTION SYNTHESIS USING PRINCIPAL COMPONENT ANALYSIS", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 1, Issue 9, pp. 2111-2116, November 2013.

Disha George is M.tech. Scholar at Raipur Institute of Technology, Raipur, Chhattisgarh , affiliated to Chhattisgarh Swami Vivekanand Technical University, Bhilai (C.G.), India.. She is an Assistant Professor MM college of Technology, Raipur, Chhattisgarh. India. Her area of interest includes Image Processing, Pattern recognition.

Yogesh Rathore received M. Tech. degree in Computer Science Engineering from Chhattisgarh Swami Vivekanand Technical University, Bhilai ,India in the year 2010.Since year 2006,he is working with the Department of Computer Science Engineering, Raipur Institute of Technology, Raipur(C.G) India affiliated to the Chhattisgarh Swami Vivekanand Technical University, Bhilai (C.G.), India. His interests include pattern recognition, image processing, neural networks, machine learning, and artificial intelligence.